Language Proficiency Level Prediction in Foreign Language Learners Texts

Tatiana Belova Daria Galimzianova

June 2022

Abstract

In the following paper we investigate the way to automatically assess essays and predict the language proficiency level. We improve English proficiency level predictions on two types of essays produced by Russian-speaking HSE students, especially among levels B1 and B2. We use the dataset with 3692 labelled essays and 63 features, including mistakes, for training CatBoost Classifier which outperforms the previous model. We also develop an assessment system for Russian as a Foreign Language text evaluation based on text complexity, which includes morphological, lexical and syntactic features. We use a model based on UDPipe Russian-SynTagRus model, complexity parameters from INSPECTOR tool paper, TORFL vocabulary minimals and a spellchecker to evaluate student's essays.

1 Introduction

Foreign language learners often have to take language tests. Giving objective evaluation to a student's writing assignments is an integral part of testing that offers a language proficiency level as a result. There are multiple ways to assess a student's performance on a written task. This may include an assessor's feedback on errors and structure, vocabulary and grammar level, grading on a special scale or evaluating on compliance with the criteria.

The scale that covers all varieties of pieces of writing produced by learners of foreign languages was introduced by the Council of Europe in 2001. [1] It is widely used as an evaluation tool for English learners around the world. Common European Framework of Reference for Languages, often abbreviated as CEFR, offers a six-level scale (A1, A2, B1, B2, C1 and C2) and descriptors for each level. The descriptors, however, are phrased in a quite obscure manner, which makes it hard to apply machine learning algorithms for CEFR level identification.

There was some research towards the problem of identification text complexity parameters, which are known to play an important part in level evaluation.[2] The language proficiency level of an essay can be assigned by a human expert on the basis of the text complexity. Some research has been carried out on the problem of identification the text complexity parameters. The study [3] explores features for syntactic complexity of the texts produced by second language learners. Another research [4] describes automatic evaluation of lexical complexity parameters in the learner texts. In this research, we use the text complexity parameters identified by the Inspector tool, since it covers several types of parameters, including morphological and discursive. Inspector is an instrument to measure the text complexity and to accentuate relevant linguistic features of the text. [5]

2 Previous work

Automatic language proficiency level detection is a problem that has been researched with the development of machine learning models. For example, [6] paper focuses on implementation of a supervised machine learning approach according to linguistic complexity. The latter is defined as a set of certain linguistic features (lexical sophistication, syntactic complexity, cohesion and accuracy) which are used to train a logistic regression model. Similar research has been conducted by [7] where the CEFR level was predicted based on complexity contours, namely syntactic and lexical richness, usage of multi-word sequences and information-theoretic measures. Recurrent neural networks proved to be effective in assigning CEFR labels in this work.

A special tool for measuring text complexity parameters has been introduced in [5] that allows to extract 63 lexical, syntactic, morphological and discursive features from any text. The Inspector tool is an open-source project.

Mistakes are not as common as complexity parameters in CEFR prediction problem, however for our task there was a wide taxonomy of errors developed by HSE School of Linguistics. Since 2012 professors and students of the HSE School of Linguistics have been working on a corpus that contains essays written by students. [8]

2.1 English as a Foreign Language

In the Russian Error-Annotated Learner English Corpus (REALEC) the essays are divided into two types: a graph description and an argumentative essay. Alongside the essays, there is annotation for mistakes. [9]

The mistakes have been manually annotated by experts. The taxonomy of mistakes is quite extensive accounting for 54 types in total. In our research we have used the texts from years 2014-2019 and their error annotations, along with Inspector complexity parameters.

2.2 Russian as a Foreign Language

An online-tool "Textometr" [10] provides text evaluation based on CEFR level, lists of optimal candidates for keywords, statistics on lexical minimals of the TORFL and lists of frequency words of the Russian language, lexical diversity measures of the text and other key statistics from any given text in Russian. This information is useful for adapting the text to educational task.

The automated assessment here is based on a regression model, trained on the dataset of more than 800 texts from Russian textbooks for foreigners, applying several machine learning and natural language processing methods.

Lexical features of text complexity are described in the article [11] on study that examines comparative complexity of Social science texts used in Russian secondary and high schools. This ongoing project investigates academic text features indicative

of its complexity at different grade levels. Based on the metrics of ten descriptive and four lexical features assessed for seven classroom textbooks authors claim that lexical diversity, frequency, abstractness and the number of terminological units are statistically significant predictors of text complexity.

3 Data Description

3.1 English as a Foreign Language

Our dataset comprises data of 3692 essays written by HSE students as an answer to the two types of questions in the Integrative English exam. This exam is held by HSE every year to test the language ability of the students. The writing tasks of this exam include two types of essays: a report on a chart and a response discussing a problem or an opinion. Every essay has 63 features, including the CEFR level label predictions, text complexity parameters identified by Inspector and values of mistakes the learner has made in the essay. The dataset contains CEFR predictions from three different neural networks with no open source code and no information about their training. The predictions were obtained from three different websites (Duolingo, Write&Improve and Grammarly) that allow the user to input a text and return a CEFR level identified by a neural network. For each essay, we only have one corresponding level label which has been chosen with an algorithm presented by O.Vinogradova and M.Bocharova at the Eurocall-2021 conference ("An experiment to obtain automated prediction of the CEFR level for learner texts of the REALEC corpus – May 2020"). In the later stages of our research, it has been decided to add more features to the dataset – the values of mistakes made in every essay.

We have trained several classifiers to predict the CEFR levels on the basis of complexity parameters alone, complexity parameters and mistakes values. In addition to this, we have manually assigned CEFR levels to 110 essays of each type and compared the resulting classifier accuracy with the one that had been previously obtained with automatic levels. As expected, manual annotation has brought the highest accuracy. This has lead to the conclusion that both complexity parameters and mistakes influence the CEFR level assignment choice. Furthermore, it is only the manual expert annotation that can lead to the significant increase in accuracy when assigning a CEFR level to an essay.

The essay dataset that we have used for training machine learning models has CEFR level annotations as prediction labels. The CEFR labels had been obtained from predictions of three neural network-driven online tools with free access: Duolingo, Write&Improve and Grammarly. We investigate the correlation between text complexity parameters, text error and CEFR levels. We focus on levels B1 and B2 as the baseline accuracy was the lowest between for the two (0.39).

3.2 Russian as a Foreign Language

We use a dataset of 93 pre-assigned essays from applicants for whom Russian is non-native language. These text are self-presentation and they are also a part of entrance exam which held to test foreign applicants' language ability. The writing task was evaluated by independent experts with six kinds of grades from level 1 (CEFR A1) to level 6 (CEFR C2) with lengths of essays vary from 50 words to over 250 words.

4 Parameters Description

4.1 English as a Foreign Language

The CEFR scale has been developed to measure the language proficiency level in foreign language learners across many European languages. The descriptors of each level in this scale constitute a list of competences that a learner possesses at each level. Although this list is comprehensible by any experienced second language teacher, it has become a challenge to quantify the levels to make them machine-readable. This is why most of the attempts to predict the language proficiency investigate the text complexity parameters. We ground our research on the complexity features obtained through the Inspector tool. The features include the following types:

1. Lexical complexity:
    - Diversity
    - Density
    - Sophistication
2. Syntactic complexity:
    - Depth of the sentence
    - Number of clauses
    - Syntactic diversity
3. Morphological complexity:
    - Derivational features
    - Inflectional features
4. Discursive Complexity:
    - Discourse-Organizing Nouns
    - Functional n-grams

The dataset includes 63 text complexity features in total, most of which have normalized values. Apart from these features, the mistakes data is also present in the dataset.

The mistakes are evaluated in absolute numbers. As a part of data pre-processing we have normalized the mistakes values, as the lengths of essays vary greatly from 100 words to over 300 words.

4.2 Russian as a Foreign Language

We do not use all of 63 foregoing complexity features and mistakes annotation from INSPECTOR tool for part of research connected with automatic assessment of Russian as a Foreign Language learners' essays as some of them were created especially for texts written in English. Some of the features were adopted to Russian language grammar rules and some of them were deleted from our tool as they were not suitable for Russian.

In addition to complexity features we use five-level scale vocabulary minimals (A1, A2, B1, B2 and C1) from TORFL and a spellchecker.

5 Hypotheses

5.1 English as a Foreign Language

We have formulated two hypotheses:

1. The complexity parameters alone are not sufficient to assign a certain CEFR level to a text. Another factor influencing the choice of a level is the errors the author of the text has made.

2. CEFR labels predicted by the neural networks are not highly accurate since we have no information about how the neural networks have been trained. Thus, manual assignment performed by the expert familiar with both the CEFR scale and the types of essays will result in a higher accuracy score.

5.2 Russian as a Foreign Language

For our baseline there was a hypothesis that the higher the arithmetic mean of numerical values of complexity parameters is, the higher learner's proficiency level is.

6 Methods

Predicting a CEFR level of an essay is a multi-class classification task, since one label for each essay is assigned. This research has the aim of improving accuracy between B1 and B2 levels for English essays and of creating similar to Inspector tool for Russian essays assignment.

Classifier	Accuracy Score
Multi-layer Perceptron Classifier	0.26
Voting Classifier (Gaussian + LogReg + Random Forest)	0.27
Gradient Boosting Classification Tree	0.32
Extra Trees Classifier	0.37
K-Nearest neighbors classifier	0.38
CatBoost Classifier	**0.42**

Table 1: Classifiers that have been trained on Inspector text complexity parameters to predict CEFR levels

	Accuracy Score	Accuracy Score after Removing Weak Features
Task 1	0.44	0.47
Task 2	0.40	0.43

Table 2: CatBoost Classifier accuracy scores before and after removing features that have the lowest importance scores

6.1 English as a Foreign Language

Initially, we have tried to predict the label through the use of different methods that are provided in the scikit-learn Python package. Namely, we have trained and tested the classifiers listed in the Table 1.

As it is seen from the Table 1, the resulting accuracy of every classifier is lower than the baseline accuracy, except the CatBoost classifier with the score of 0.42. The further classification involved a series of experiments with essays written for task 1 and task 2 separately. The essays of the two types have to be analyzed in isolation because they serve different communicative goals, hence the vocabulary and the text structure are distinct from each other. Table 2 offers the accuracy scores for both types of tasks. Scores of 0.44 and 0.40 that we achieved classifying both types separately still seemed to be quite low. So, a set of experiments on removing weak features were conducted. To decide whether a particular feature was important or weak, we used the classifier method of calculating the feature importance. It shows how much on average the prediction changes if the feature value changes. Then, we heuristically removed features that have shown an importance of 0.1-1 (the biggest value in the table being 9). The highest score we could achieve by removing different unimportant features is presented in the Table 2.

The experiments described above have only used Inspector text complexity parameters as features and CEFR level as prediction labels. Moreover, the labels had been previously predicted by a combination of neural networks (Duolingo, Write&Improve and Grammarly).

To test our first hypothesis, the mistakes features has been added to the dataset. Then the absolute values of mistakes have been normalized by the length of each

	Complexity Parameters + Normalized Errors	Complexity Parameters + Normalized Errors + Manual Labels
Task 1	0.57	0.64
Task 2	0.48	0.76

Table 3: Accuracy scores for CatBoost classifier trained on complexity parameters and error values to predict automatically or manually assigned labels

essay. The accuracy score has improved, especially for the first type of essay (graph description), which can be seen from the Table 3. The accuracy for the second type has increased as well but still has not reached the appropriate levels. We have trained a classifier on both complexity parameters and mistakes values, which has resulted in improved accuracy for both types of essays, consequently the first hypothesis is confirmed.

The second hypothesis requires an human-annotated data. We have used the detailed CEFR Descriptors (2020) [1] to manually assign CEFR levels to 110 essays of each type.

We used the manually annotated labels, complexity parameters and normalized mistakes to train the classifier with 80/20 training and testing sets. This has brought a substantial increase in accuracy, which can be observed from the Table 3. This proves the second hypothesis.

6.2 Russian as a Foreign Language

For the second part of our research connected with Russian essays assignment we created a model which is based on UDPipe 2.5 Russian-SynTagRus model [12]. It tokenizes, lemmatizes text, does a POS-tagging, divides text into sentences and saves it in conllu format for further parsing. The second part of our tool – Parser itself – uses complexity parameters from INSPECTOR tool paper, TORFL vocabulary minimals and a spellchecker from Python 3 library "pyspellchecker" with additional Russian word frequency list of 1.5 million word forms.

Unlike Inspector tool we decided not to use machine learning methods for our baseline, and used statistic method instead. According to [5] learner's proficiency depends on number of various vocabulary means, dependent clauses, sentences, words before the root of the sentence; high indicator of lexical density; length of sentences. Learners of higher levels use more sophisticated vocabulary, their texts are usually longer, syntax and vocabulary are more complicated.

We used our tool on texts from dataset with 93 essays, calculated the agreement between tool's and annotator's grades - the result showed 85% of agreement between them.

We also analyzed vocabulary of the text using TORFL minimals, and displayed the percentage of vocabulary of each level.

Even though we had a hypothesis that high-level essays would perform high arithmetical mean, it turned out that number of words and number of lemmas affected the result. Lower-level students' essays with large number of words were given the higher level by baseline than human-annotator did.

Consequently, we decided to train a simple Logistic Regression classifier which takes vectorized values of complexity parameters of the essay and predicts its level.

Mistakes, if there are any in the text, are displayed separately and they are not used while calculating arithmetic mean, so they do not affect the final grade.

Our tool also categorizes words from essay by their level using TORFL minimals and displays how many words are in each category. If the word does not fit any category, it is labelled with "Not in lists".

7 Conclusion

Foreign language learners produce a variety of writing pieces which are not always easy to assess. One of the most popular tools in evaluating L2 learners is the CEFR scale. The descriptors for each level are quite extensive and easy to use for a trained professional. However, there are little to no quantifiers present in the descriptors, which makes it hard to train a machine learning model to predict a CEFR level of a text produced by an L2 learner.

In this research, we have experimented with different parameters and have reached the conclusion that both the complexity and errors influence the CEFR level of a text.

It has been also observed that the accuracy of the labels produced by the neural networks trained on general English texts is lower for the texts of academic English we have dealt with. However, we have manually annotated only a small part of the dataset (220 essays in total), which is not enough for accurate predictions. There was only one annotator, which is another drawback of this study. Further research might include more annotations performed by several experts.

For the Russian language, we presented a tool for automatic essay evaluation that is based on our research. Our tool is aimed at helping Russian teachers and professors to check a great amount of essays in short time. The tool can be found at https://github.com/tatiana-belova/russian_tool.

Whilst our tool allows essay assignment, there are several limitations. Errors analysis is an integral part of text evaluation, and mistakes affect the final grade, so there should be full error classification, which takes different kinds of mistakes like lexical, morphological, syntactic and discursive into consideration while evaluating essay. Also, text logic is area which is hard to quantify: a text might contain some upper-level vocabulary and grammar but the whole structure of the essay could be nonsense and illogical. This is true for both English and Russian learner essays. Future work may consider the relationship between vocabulary sophistication and syntactic complexity.

References

[1] Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press, 2001.

[2] Automatic approach to text difficulty measurement for rfl.

[3] Xiaofei Lu. Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496, 2010.

[4] Kenneth Hyltenstam. Lexical characteristics of near-native second-language learners of swedish. *Journal of Multilingual & Multicultural Development*, 9(1-2):67–84, 1988.

[5] Irina M Panteleeva, Olga N Lyashevskaya, and Olga I Vinogradova. Inspector: The tool for automated assessment of learner text complexity. *Higher School of Economics Research Paper No. WP BRP*, 79, 2019.

[6] Thomas Gaillat, Andrew Simpkin, Nicolas Ballier, Bernardo Stearns, Annanda Sousa, Manon Bouyé, and Manel Zarrouk. Predicting cefr levels in learners of english: the use of microsystem criterial features in a machine learning approach. *ReCALL*, 34(2):130–146, 2022.

[7] Elma Kerz, Daniel Wiechmann, Yu Qiao, Emma Tseng, and Marcus Ströbel. Automated classification of written proficiency levels on the cefr-scale through complexity contours and rnns. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 199–209, 2021.

[8] Elizaveta Kuzmenko and Andrey Kutuzov. Russian error-annotated learner english corpus: a tool for computer-assisted language learning. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 87–97, 2014.

[9] Andrey Kutuzov and Elizaveta Kuzmenko. Semi-automated typical error annotation for learner english essays: Integrating frameworks. In *Proceedings of the fourth workshop on NLP for computer-assisted language learning*, pages 35–41, 2015.

[10] Lebedeva Maria Y. Laposhina, Antonina N. Textometr: an online tool for automated complexity level assessment of texts for russian language learners. *Russian Language Studies*, 19(3):331–345, 2021.

[11] Anna Churunina, Marina Solnyshkina, Elzara Gizzatullina-Gafiatova, and Artem Zaikin. Lexical features of text complexity: the case of russian academic texts. *SHS Web of Conferences*, 88:01009, 01 2020.

[12] Milan Straka. CoNLL 2017 shared task - UDPipe baseline models and supplementary materials, 2017. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.